Supporting Information Appendix for:

The Persistence of Racial Discrimination: A Meta-Analysis of Field Experiments in Hiring

Lincoln Quillian, Devah Pager, Ole Hexel,
Arnfinn H. Midtbøen

correspondence to: l-quillian@northwestern.edu

Includes:

Supporting Text
Figs. S1 to S2
Tables S1 to S10

Supporting Information Appendix

1. Problems in Assessing Trends in Discrimination

Methods for measuring discrimination are notoriously flawed, a problem only compounded by attempts to make comparisons over time. The most common approach to studying trends in racial discrimination has been the residual method: based on a statistical model of an outcome, the residual race gap left unexplained after other factors are accounted for is attributed to discrimination. If this residual is reduced over time, discrimination is thought to have declined. This approach suffers from the significant weakness that its validity rests crucially on controlling for all other factors that influence the outcome and may vary between racial groups – a circumstance that is generally impossible to verify (1, 2). This problem is compounded by attempts to draw comparisons over time, as the source and importance of relevant controls (or unobservables) may shift over time.

Other approaches to assessing trends in discrimination have relied on survey and institutional reports. These methods have different but no less serious problems. One approach relies on self-reports of discrimination from targets (e.g. 3). The weakness of this method is that it cannot detect discrimination that targets are not aware of; and, conversely, targets may sometimes mistakenly attribute a poor outcome to racial discrimination when the outcome has a different basis. A second method uses the frequency of formal complaints of discrimination from targets or lawsuits alleging discrimination (4). This method too only captures discrimination that victims are aware of, and formal complaints or lawsuits are strongly influenced by institutional factors that discourage or encourage reporting or lawsuits (5).

A final method relies on interviews with potential perpetrators, but this approach faces obvious potential problems with underreporting of socially unacceptable conduct, and also cannot capture discrimination grounded in implicit or subtle attitudes that perpetrators may not be aware of. Again, we face the problem that these analyses require strong assumptions that cannot be tested, leading to unresolved questions of whether discrimination is really changing or whether apparent changes are artifacts in measurement (1, 6, 7).

In an effort to address the problems of measurement and potential omitted variable bias plaguing research in this area, researchers have increasingly turned to experimental methods (8). Field experiments, in particular, offer a powerful design for isolating the causal effect of discrimination within the context of real-world hiring decisions. Researchers conducting field experiments have a high degree of control over the experimental conditions (whether matched testers or racially-identifiable names randomly assigned to resumes), providing a strong basis from which to draw causal conclusions about hiring discrimination. Likewise, situating these experiments in the context of actual hiring decisions offers conclusions that readily generalize to real labor market dynamics.

That being said, field experiments of discrimination have also been subject to important critique. Field experiments typically focus on a single skill level (or a narrow range) for their tests of discrimination, thereby potentially missing variation in rates of discrimination facing those of greater of lesser skill levels and across a wider range of occupations. In-person audits can suffer from spurious effects due to the poor matching of test partners and/or experimenter effects due to the expectations of testers (9). Resume audits, which rely on racially-identifiable names for their key experimental treatment, may confound racial discrimination with effects driven by class, by relying on names that may signal both (10). Both types of audit studies typically rely on random samples of job listings for their tests of discrimination. Selective applications by job seekers and selective recruitment by employers may complicate the degree to which estimates of discrimination from audit studies map on to real-world experiences with discrimination (11, 12). These potential limitations are by now well known and have received extensive attention in the research literature. While by no means perfect, field experiments continue to be considered the most valid measure of hiring discrimination available (1). In the current study, we capitalize on the meta-analysis approach which draws information from across studies. We can therefore explicitly model design issues (such as resume-based versus in-person) and variability across research studies.

2. Estimates of Average Discrimination Levels

The first step of our analysis is to consider the overall levels of discrimination by group and the extent of heterogeneity across studies. Figure S1 shows a forest plot of the discrimination ratios of 24 studies in which the target group is African-Americans, contrasted to whites.

In Figure S1 an overall average discrimination ratio for 1990 to 2015 based on random-effects meta-analysis is shown as the final diamond on the table (see Methods and Materials and the SI Appendix section 9 for details of the meta-analysis model). The results indicate that on average whites receive 36% more positive responses to job applications than African-Americans. A 95% confidence interval for the effect is 25% to 47% more callbacks. This reinforces the conclusion of many in-person and resume audits regarding the persistence of hiring discrimination against African-Americans, and provides a broader overall estimate of the average prevalence of discrimination in hiring by combining information from 24 studies. If we use all data from 1972 to 2015, a total of 24 estimates from 24 studies, whites receive on average 34% more positive responses than African-Americans, with a 95% CI of 23% to 46% more.

What accounts for the variability in estimates across studies? The model estimates that 67.3% of variability (I-squared) reflects differences resulting from differences in study characteristics (e.g. year, applicant education level, in-person vs resume audit, etc.). The remaining 32.7% of variability between studies could be accounted for by random variation in outcomes of individual studies. A significance test strongly rejects the hypothesis that the between-study variability is zero ($p<.001$), supporting a random-effects specification.

Figure S2 provides a forest plot for studies estimating rates of discrimination against Latinos relative to whites. On average whites received 24% more positive responses than Latinos, with a 95% confidence interval of 15% to 33% more.

3. Discrimination Trends

Table S3 presents estimates of the random-effect meta-regression used to create the trend line estimates and weights in Figures 1 and 2. In the model the discrimination ratio is logged. The coefficient of year in table S3 may be interpreted as percentage change in the discrimination ratio each year. For instance the coefficient of .004 for 1990-2015 indicates a trend upward of

.4% in the discrimination ratio per year. Both lines slope upward (coefficients of year in table S3 of greater than zero).

The rightmost numeric column of table S3 shows the coefficient that is the basis for the line of best fit in figure 2. Here the coefficient is less than zero, indicating the downward slope in discrimination against Latinos shown in figure 2.

4. Models of Discrimination Trends with Controls and Sensitivity Analyses

We then performed sensitivity analyses of changes in the dependent variable and built a model with additional controls to account for characteristics of studies that may confound the time trend. Because there are too few studies for Hispanics to support controls, we only perform the multivariate analyses for African-Americans. Model estimates are shown in table S4.

Coefficients of the model are shown in table S4. The coefficients for year are shown with confidence interval lines in figure 3. Models 2 to 4 alter the dependent variable (models 2 and 3) or sample (model 4). The results are robust to this change: we never have a statistically significant downward slope, and in all but one modification the point estimate is still an upward trend.

Models 5 to 8 add controls for characteristics of applicants and studies. These are added to control for characteristics of studies that could confound the trend over time and might influence discrimination levels. We include as controls common covariates that have been used to model discrimination rates in field experimental studies (e.g. 13, 14) and studies suggested by theoretical accounts of factors that might influence discrimination (e.g. 15, 6, 7). This includes study method, gender of the applicants, education level of applicants, presence of a fictional criminal background, metropolitan unemployment rates, region, and occupational sectors. Past theoretical literature has suggested reasons each of these may influence discrimination rates; for instance when unemployment rates are high, employers may be more likely to indulge discriminatory tastes given the wide range of applicants they have to choose from. However, we find that few of these controls are statistically significant predictors of discrimination.

Model 8 trims the model by dropping variables that consistently had t-ratios below 1.5. None of the controls has much effect on the time trend in discrimination, which remains insignificant and usually slightly upward in direction for white vs. African-American (indicated by the positive coefficient).

Alterations of the dependent variables for the Latino models are shown in table S5. These results show a bit less evidence for downward trend than for the model in table S3. The number of studies with Latino respondents is too low to allow for additional covariates.

In table S6 we use an alternative procedure to compute discrimination trend estimates: a model that pools African-American and Latino effect sizes. We use African-American as the base trend and allow an interaction of Latino by year. Coefficients of controls are constrained to be similar for the two groups.

Six of the studies in the analysis provide estimates of discrimination against both African-Americans and Latinos, which are represented as separate effects. We account for the clustering of the estimates within study by using robust standard errors allowing for correlated effects within study and a small sample adjustment (procedures discussed in 16). Model estimates are based on assumed correlation of rho=.8. Sensitivity analysis showed only slight changes in estimates with different values of rho.

The first two models show results including the linear trend, and dummies for whether the effect size is for the Latino target group (with the black target group as the reference). The coefficient for the Latino effect suggests less discrimination against Latinos, although this difference is only statistically significant (at $p<.1$) in the initial two models. Like the models estimated only for African-Americans, the pooled estimates provide no evidence of downward trend: all point estimates except one show a positive trend, and the one point estimate with negative trend is almost flat (model 4, year coefficient=-.004).

5. Publication Bias

Is it possible that recent studies finding little or no discrimination are more likely to remain unfinished or unpublished, thus excluded from our analysis and biasing our estimates? This problem, known as publication bias, represents a possible threat to the validity of any meta-analysis (17).

We find some evidence of publication bias in the analysis presented in table S7. Studies published in academic journals show somewhat higher discrimination ratios than those published elsewhere (reports, working papers, etc.), a result that is marginally statistically significant ($p<.1$). Because we were aware of this potential issue, we went to great lengths to include unpublished studies and reports. By contacting all known authors of audit studies and other researchers who specialize in the study of discrimination, we attempted to learn of any past, recent, or ongoing study that would not show up in our bibliographic search. Indeed, 12 of the 31 discrimination estimates in our study come from sources that were unpublished when initially included in our analysis. We believe, then, that any existing publication bias is unlikely to have affected our estimates drawing from both published and unpublished sources.

While this reduces the danger from selection into formal publication in academic journals or books, if studies that do not find discrimination are less likely to be written up at all—as working papers, reports, journal articles, etc.—there still may be an important form of publication bias even when unpublished studies are included ("write-up bias" may be a more accurate term for this).

As a more formal investigation into the possibility of publication or write-up bias (both of which create the problem that only studies finding race differences are likely to end up producing a report that can be included in our analysis), we included a study-level predictor that is diagnostic these problems. Specifically, we created a dummy variable to indicate whether the study design implied a primary focus on race or some other attribute. For instance, some studies focused on effects of particular educational qualifications, labor market histories, or a criminal record on receiving a positive response from an employer. While these studies also included racially diverse applicant profiles, allowing for an estimate of the effects of race on hiring outcomes, the primary emphasis of the study was on something other than racial discrimination. Note that in many of these studies, the variable of interest is included as a within-pairs contrast while secondary variables (i.e., race) are included as a between-pairs contrast. Given the way these priorities are expressed in the design of the study, the question of whether or not a study has race as its primary focus is not simply a matter of post-hoc interpretation. If write-up bias or publication bias is a serious problem, then we should find that race-focused studies tend to find more racial discrimination than not-race-focused studies, since a significant finding on race should be more important for write-up and publication in the case of studies with race as their primary emphasis. We prefer this test to tests in the meta-analysis literature based on symmetric distributions of effects, because these tests require the assumption that other confounding variables are not related to study size or effect sizes (see 17, section 23.3.1).

Results are shown in table S7. We find on average effect sizes in not-race-focused studies are somewhat larger than effect sizes of studies focused on race (not a statistically significant difference), as indicated by the coefficient above one, opposite the direction that publication or write-up bias would predict. According to this test, publication or write-up bias is unlikely to have produced inflated discrimination estimates.

6. Excluded studies.

Overall our search located 34 studies that were U.S.-based field experiments of hiring and included contrasts between white and non-white applicant profiles who were on-average equivalent in their labor-market relevant characteristics (e.g. education, experience level in the labor market, etc.). Two studies were excluded because it was not clear if employers were the ones making decisions producing discrepant outcomes because applications were conducted through an employment agency. One study contrasted whites to Arab-Americans; we excluded this study since it was the only study with this target group. One study did not report counts and the authors declined our request for counts of outcome by target groups. Two studies were excluded because they used mixed non-white groups and did not break out results separately for African-American and Latino applicants. All other studies focused on whites contrasted to African-Americans or Latinos (or both). Our remaining 28 studies yielded 24 estimates of discrimination against African-Americans and 9 against Latinos relative to whites. For most analyses in this paper, we exclude studies before 1989, which leaves us with 21 estimates of discrimination against African-Americans and 9 against Latinos from 24 studies (six studies include estimates of discrimination against both African-Americans and Latinos).

7. Coding

To ensure reliability, each study was coded independently by two raters. The first rating was completed by the third author under direction of the first author of this paper. The second rating was performed by two undergraduate students who were hired to conduct a second coding using the rubric. We then reconciled the results of the two codings, performing further investigation to find the correct answer on coding decisions in cases of disagreement. The variables coded were factual in nature (e.g. year of publication, counts of positive and negative responses for the white and non-white group, etc.); the main sources of disagreement in coding were difficulty in understanding the text or procedures of a particular study, or occasional judgment calls about what “fit” on a particular category. For instance, such judgments include decisions about whether working in a warehouse stockroom counts as “blue collar” employment (we did code it as blue collar), or whether an employer’s response that they would keep an applicant’s resume on file and might eventually request an interview constitutes a callback (we did not count this as a callback). In cases of disagreement or high uncertainty in the reconciliation process, the first author examined the study and broke the tie by assigning a code.

The coding involved two levels of information: study level and effect level. Study level characteristics are constant for the entire study, such as year of publication and type of publication outlet (academic journal, government report, etc.). Effect estimates refer to estimates of discrimination against a non-white group, with the number of effect sizes for a study depending on the number of target groups the study includes. A study that contrasts both African-Americans and Latinos to whites would produce two effect sizes.

We coded effects that measure discrimination based on counts of hiring outcomes by racial or ethnic group. Most studies included this information in the write-up. When the study did not include counts of outcomes in their research report, we requested counts from the authors, which we received for all studies with black or Latino target groups. We used all white and minority applicant profiles in computing discrimination ratios, except for cases in which the groups were non-equivalent in their labor market characteristics (most often where minorities were given somewhat stronger background qualifications than whites). For instance, one study included contrasts between white applicants with criminal records and black and Latino applicants without criminal records (19), and we do not include these audits because of the non-equivalence of the characteristics of white and minority testers (this study also included some audits between equivalent groups and is thus not excluded entirely). In our baseline analysis we include applicant profiles with characteristics such as a criminal background or a disability as long as this condition was equally present between white and minority testers. We perform some sensitivity analysis to illustrate changes in results when perhaps atypical groups like the disabled or those with a criminal background are eliminated (see SI Appendix section 4 and table S4).

We excluded some audits that were part of the New York Audit study reported in (14, 18) because they were based on between-pair comparisons, when within-pair comparisons focused on race were available from the same audit study. Adding in these audits has no effect on our results.

8. The Discrimination Ratio contrasted to the Difference in Proportions or the Log Odds Ratio

We use the ratio of the proportion of callbacks received by white applicants to the ratio received by nonwhite applicants to measure discrimination. Two other candidate measures that could be used instead are the difference in proportions or the odds ratio (19).

The difference in proportions is a measure widely used in the correspondence and audit literature. In our context this measure is the difference between the percentage of callbacks received by whites and the percentage of callbacks received by minorities ($\frac{c^w}{n^w} - \frac{c^m}{n^m}$). This is a logical measure for single studies, but we think it is less well-suited for a meta-analytic context in which base rates vary over studies. When the base rate of callbacks gets relatively close to 0% (many studies have low callback rates), then the floor places a limit on the size of the difference in proportions. When base rates differ over studies, the floor places varying limits across studies.

This problem is especially clear in paired field experiments. In paired studies one or more majority and minority applicants apply for the same job. In most field experiments the most common outcome is that neither the majority nor the minority auditor gets a callback for an interview. This "neither" outcome provides no information about discrimination. But the frequency of this outcome sets an upper limit on the size of a difference ratio of discrimination. For instance, if in 80% of audits neither applicant gets a callback, then the maximum difference in the proportion of callbacks between groups is .2. By contrast, ratio measures are not automatically limited by low base-rates of callbacks.

The result of using a difference measure in the presence of unequal base rates is greater heterogeneity in the outcome across studies in the meta-analysis, because the difference measure absorbs base rate variability. This is a problem that has been recognized in methodological discussions in the meta-analysis literature as an undesirable feature of the difference in proportions as an outcome measure (see section 13.2.1 in 19).

We also find the ratio measure is preferable because the base rate of callbacks must implicitly be invoked to understand the implications of differences in callback rates for racial disparities in hiring. If whites receive 8% callbacks and African-Americans receive 4% callbacks, then whites receive 200% as many callbacks per application submitted, an outcome strongly favoring whites in hiring. By contrast, if whites receive 44% callbacks and African-Americans receive 40% callbacks, whites receive 10% more callbacks per application submitted, a much smaller advantage in getting a job. Using the difference measure in both cases the racial disparity is measured to be the same at 4%.

Because the difference in proportions is a widely used measure in the field experimental literature, we conducted a sensitivity analysis using this measure. Results using the difference in proportion measure for African-Americans are shown in table S8, column 1. We use the applicant attributes model, which we take as our most basic model. The point estimate for the time trend in this model is upward, and the coefficient is small and not statistically significant – similar to what we find with the ratio measure.

Result for the difference measure for Latinos are shown in column 2 of table S8. There is a statistically significant downward trend with this measure.

A second candidate measure is the natural log of the odds ratio. The odds ratio measure takes the ratio in the odds of the outcome between groups in place of the ratio in the proportions. Formally, if c^w is the number of callbacks received by whites, and c^m is the number of callbacks received by blacks or Latinos, and n^w is the number of applications submitted by white applicants, and n^m is the number of applications submitted by black or Latino applicants, then the odds ratio is $\frac{\frac{c^w}{n^w-c^w}}{\frac{c^m}{n^m-c^m}}$. The odds ratio is logged to make its distribution more symmetric before regression analysis. This measure has good statistical properties, but we prefer the ratio of proportions to the odds ratio because of the greater interpretability of the ratio of proportions. Ratios of proportions are more intuitive than ratios of odds. Further, many field experimental studies also use ratio of proportion as their basic measure of discrimination (e.g. 14), but we know of none that use odds ratios for basic group comparison of outcomes.

Results using the log odds ratio as the outcome for white vs. African-Americans are shown in table S9 column 1. Again, the slope of the year variable is slightly upward and very close to zero. The main result is unchanged from the log ratio of proportions. Results for Latinos are shown in table S9 column 2.

For Latinos, we find a statistically significant downward trend with both the difference in proportions and the log odds ratio. This strengthens the case for decline in discrimination against Latino job seekers across the years we consider. However, our inability to include controls due to the small number of studies including Latinos weakens our ability to draw inferences about trends in anti-Latino discrimination. And modifications to the dependent variable discussed in section 4 and shown in table S5 result in non-significant year coefficients for Latinos.

9. Weighting and Model Estimation

Meta-analysis requires estimating the sampling variability of the discrimination estimate for each study. To estimate this we use standard formulas for variability of a ratio due to sampling error and the counts of outcomes from each studies. For studies that are unpaired or do not report paired outcome, the variance of the log risk ratio in the ith study is estimated by (see 20, formula 5.3):

$$\sigma_i^2 = \mathrm{Var}(\ln(y_i)) = 1/c^w_i - 1/n^w_i + 1/c^m_i - 1/n^m_i$$

Where c^w is the number of callbacks received by whites, c^m is the number of callbacks received by blacks or Latinos, and n^w is the number of applications submitted by white applicants, and n^m is the number of applications submitted by black or Latino applicants

For studies that use a paired design – with one minority and one white applicant applying for each job – and report paired outcomes, we use an alternative formula to account for the pairing. If p^a are the number of pairs in which both majority and minority testers receive a callback, p^b are the number of pairs in which the majority tester received a callback but not the minority, p^c are the number of pairs in which the minority tester received a callback but not the majority, and p^d are the number of pairs in which neither tester received a callback, then the variance of the log odds ratio in the ith study with paired data is (see 21):

$$\sigma_i^2 = \mathrm{Var}(\ln(y_i)) = \frac{p_i^a + p_i^d}{(p_i^a + p_i^b)(p_i^a + p_i^c)}$$

The meta-regression model is:

$$\ln(y_i) = x_i\beta + u_i + e_i, \text{ where } u_i \sim N(0, \tau^2) \text{ and } e_i \sim N(0, \sigma_i^2)$$

where β is a $k \times 1$ vector of coefficients (including a constant), and x_i is a $1 \times k$ vector of covariate values in study i (including a 1 for a constant).

The random effect variance (τ^2) is estimated with the parameters as part of the meta-analysis model. Effectively, the random effect is estimated by the extent of residual variation in the outcome that cannot be accounted for by sampling variability for each study (σ_i^2) or the covariates. Estimation is by restricted maximum likelihood, which has the advantage of giving nearly fully efficient estimates and nearly unbiased estimation of variance components in smaller samples. For further details of estimation, see 22.

In practice we estimated the basic meta-analyses and meta-regressions in Stata using the "metan" suite of commands (23). The Knapp-Hartung modification, which we employ in all tables except table S6 uses the t-distribution rather than the normal distribution for inference. The t distribution has been shown to provide more accurate tests and confidence intervals in simulation studies (24).

For the pooled model shown in table S6, we use the "robumeta" command in Stata written by Tipton and co-authors (16). This include robust estimation methods to adjust for correlated effect sizes in cases for which the African-Americans and Latinos effect sizes are from the same studies.

A bibliographic list of studies in the meta-analysis, data, and code are available from the discrimination meta-analysis project website at: http://sites.northwestern.edu/dmap

10. Trends in Covariates

Altonji, Elder, and Taber (25) argue that confounding by observable covariates can provide evidence about confounding by unobservable covariates, viewing the observable covariates as a random sample on a larger set of unobserved covariates. This suggests that, under assumptions they outline, we can test for potential bias in unobservables by examining the correlation of observed variables with year of survey, our primary variable.

Table S10 shows an OLS regression of year of fieldwork on all of the covariates. An F-test regressing the year of study on all of the control variables has a p-value of .1938, suggesting no clear evidence of trend in the observed covariates over years. Under the logic that Altonji et al. outline, this suggests unobservables also are unlikely to be correlated with year in a strong enough way to create substantial bias in estimates of year.

Looking at bivariate relationships to year and models with smaller sets of covariates (not shown), two variables are sometimes statistically significant in predicting survey year: audit design (in-person vs. resume audit) and the inclusion of applicants with a criminal background. These are not surprising in light of trends in job application practices and trends in incarceration: resume audits have become more popular as the rise of the internet has made online application procedures increasingly common; including testers with fake criminal records has become more popular because of the growth of incarceration and concern about its consequences.

We include sensitivity analyses in our base results excluding in-person audits (resume audit studies only) and excluding applicant profiles with fake criminal backgrounds. These results are shown in table S4, and the slope estimates shown in figure 3 in the main text. Excluding auditors with criminal records and limiting our study to only resume audit studies both produce no trend in discrimination ratios for African-Americans. We conclude that there are no trends in the covariates that appear potentially problematic for our results.

References

1. National Research Council (2004) Measuring racial discrimination. eds Blank R-M, Dabady M, Citro C-F (National Academies Press, Washington, DC).
2. Charles K-K, Guryan J (2011) Studying discrimination: Fundamental challenges and recent progress. *Annu Rev Econom* 3:479-511.
3. Feagin J-R, Sikes M-P (1994) *Living with racism: The black middle-class experience* (Beacon Press, Boston).
4. Roscigno V-J (2007) *The face of discrimination: How race and gender impact work and home lives* (Rowman & Littlefield Publishers, Lanham).
5. Nielsen L-B, Nelson R-L (2005) Scaling the pyramid: A sociolegal model of employment discrimination litigation. *Handbook of Employment Discrimination Research*, eds Nielsen L-B, Nelson R-L), pp 3-34.

6. Quillian L (2006) New approaches to understanding racial prejudice and discrimination. *Annu Rev Sociol* 32:299-328.
7. Pager D, Shepherd H (2008) The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. *Annu Rev Sociol* 34:181-209.
8. Pager D (2007) The use of field experiments for studies of employment discrimination: Contributions, critiques, and directions for the future. *Ann Am Acad Pol Soc Sci* 609:104-133.
9. Heckman J-J (1998) Detecting discrimination. *J Econ Perspect* 12(2):101-116.
10. Gaddis, S-M (in press) How black are Lakisha and Jamal? Racial perceptions from names used in correspondence audit studies. *Soc. Sci.*
11. Fix M, Struyk R-J eds (1993) *Clear and convincing evidence: Measurement of discrimination in America* (Urban Institute Press, Washington DC).
12. Pager D, Pedulla D-S (2015) Race, self-selection, and the job search process. *Am J Sociol* 120(4):1005-1054.
13. Bertrand M, Mullainathan S (2004) Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. Am Econ Rev 94(4):991-1013.
14. Pager D, Bonikowski B, Western B (2009) Discrimination in a low-wage labor market: A field experiment. Am Sociol Rev 74(5):777-799.
15. OECD (2013) Discrimination against immigrants – measurement, incidence and policy instruments. *International migration outlook 2013*, (OECD Publishing), pp 191-230. Retrieved April 1, 2017. (http://www.oecd-ilibrary.org/social-issues-migration-health/international-migration-outlook-2013/discrimination-against-immigrants-measurement-incidence-and-policy-instruments_migr_outlook-2013-7-en).
16. Tipton E (2015) Small sample adjustments for robust variance estimation with meta-regression. *Psychol Methods* 20(3):375-393.
17. Sutton A-J (2009) Publication bias. *The handbook of research synthesis and meta-analysis*, eds Cooper H-M, Hedges L-V, Valentine J-C (Russell Sage Foundation, New York), 2nd Ed.
18. Pager D, Western B, Sugie N (2009) Sequencing disadvantage: Barriers to employment facing young black and white men with criminal records. *Ann Am Acad Pol Soc Sci* 623(1):195-213.
19. Fleiss, JL, Berlin JA (2009) Effect sizes for dichotomous data. The handbook of research synthesis and meta-analysis, eds Cooper H-M, Hedges L-V, Valentine J-C (Russell Sage Foundation, New York), 2nd Ed.
20. Borenstein M, Hedges L-V, Higgins J-P-T, Rothstein H-R (2009) *Introduction to meta-analysis* (John Wiley & Sons, Chichester, U.K) p 421.
21. Zou G-Y (2007) One relative risk versus two odds ratios: Implications for meta-analyses involving paired and unpaired binary data. *Clinical Trials* 4(1):25-31.
22. Raudenbush S-W (2009) Analyzing effects sizes: Random effects coding. *The handbook of research synthesis and meta-analysis*, eds Cooper H-M, Hedges L-V, Valentine J-C (Russell Sage Foundation, New York), 2nd ed, pp 295-316.
23. Bradburn M, Deeks J, Altman D (1998) Metan – an alternative meta-analysis command. *Stata Technical Bulletin* 44: 4-15.
24. Knapp G, Hartung J (2003) Improved tests for a random-effects meta-regression with a single covariate. *Stat med* 22: 2693-2710.
25. Altonji, J, Elder T, Taber C (2005) Selection on observed and unobserved variables: assessing the effectiveness of catholic schools. *J Political Economy* 113 (1): 151-184.

Figure S1: Discrimination Ratios African-Americans

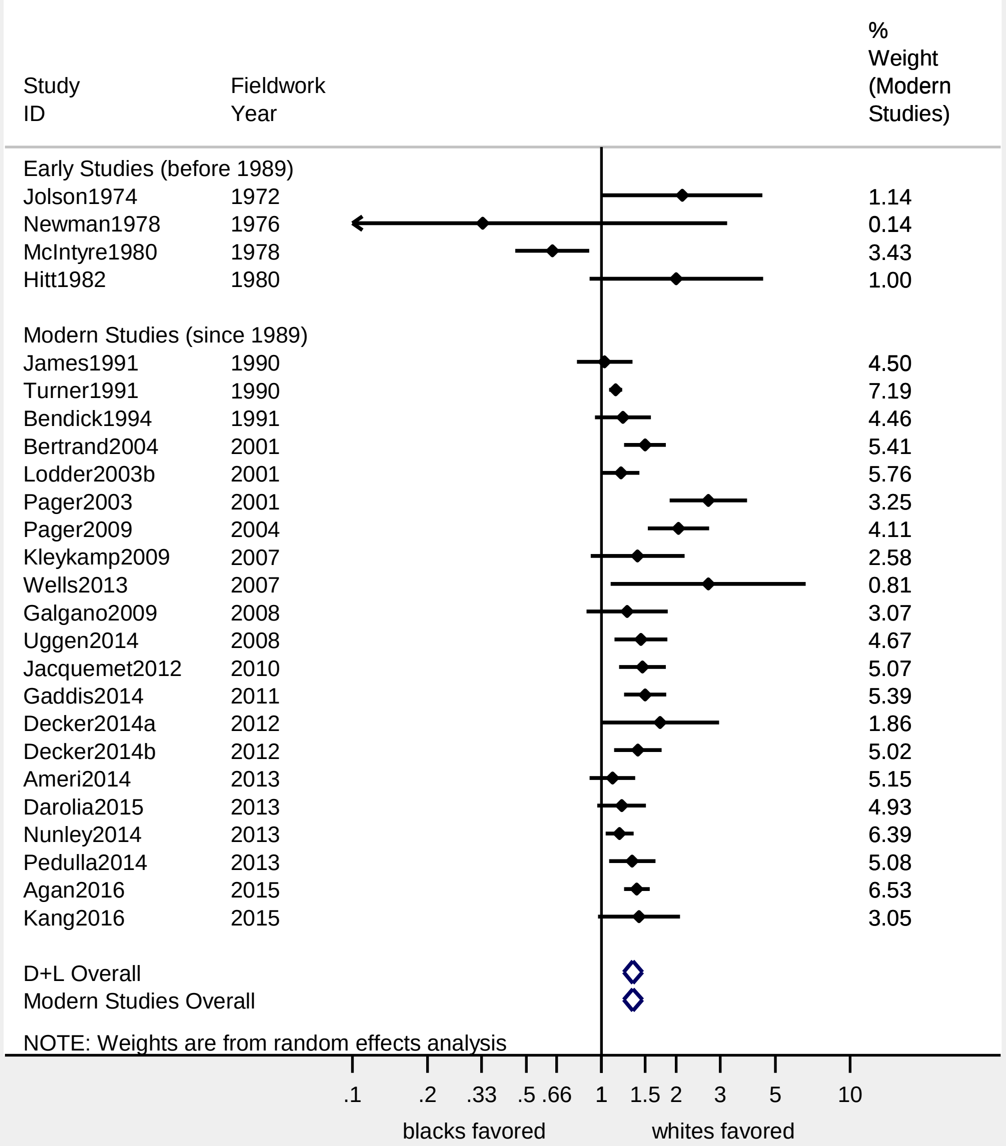

Figure S2: Discrimination Ratios, Latinos

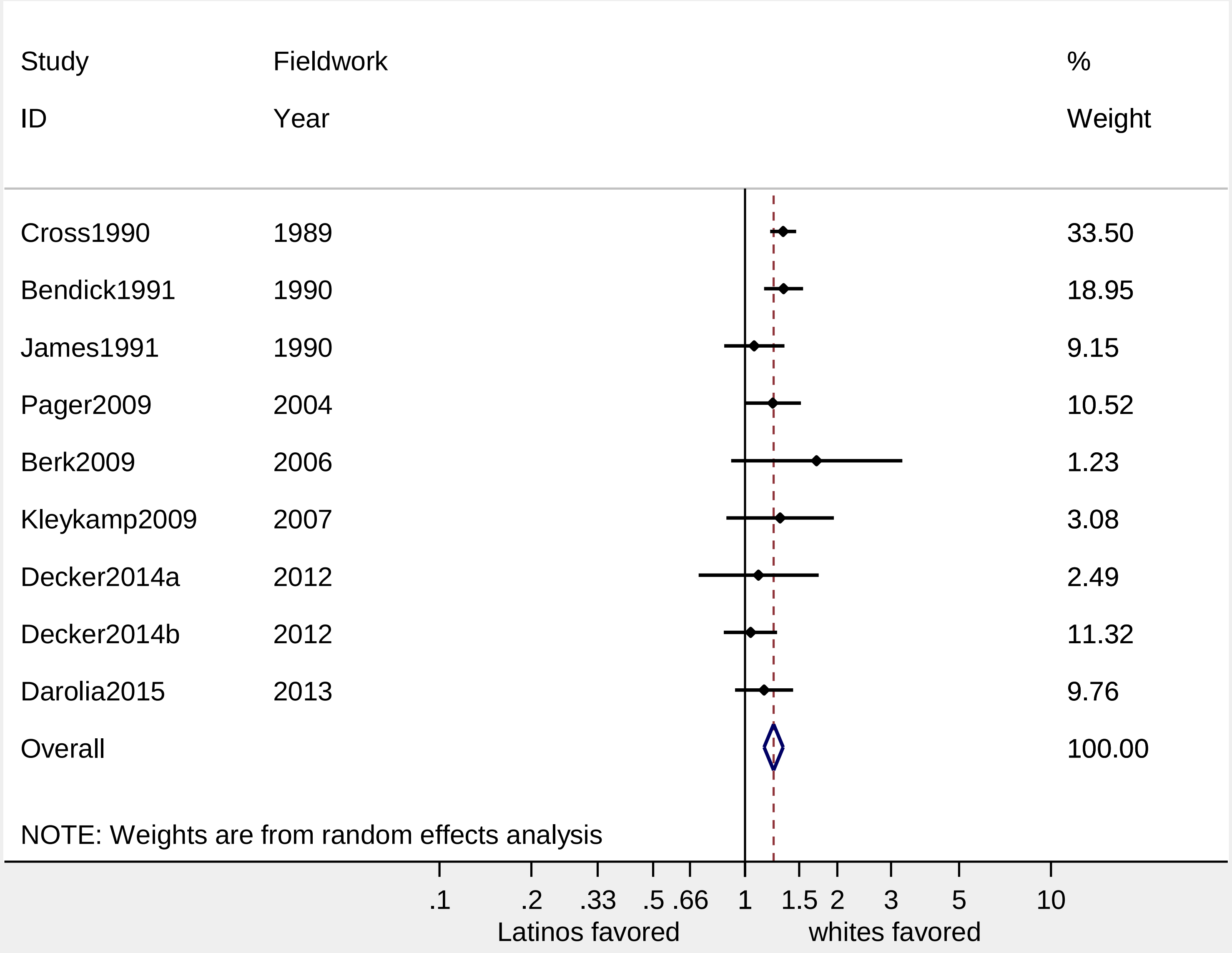

Table S1: Characteristics of Field Experiments of Discrimination Since 1989 (Categorical variables)[(a)]

	Count	Percentage
Study Method		
In-person audit	12	40%
Resume audit	18	60%
Target Group		
Black/African-American	21	70%
Hispanic/Latino	9	30%
Race is Not a Primary Focus of the Publication		
Yes	14	45%
Publication Type		
Dissertation or MA thesis	3	10%
Journal article	18	60%
Report	7	23%
Working paper	2	7%
Paired design with mixed-race pairs? [(b)]		
Yes	7	23%
No	23	66%
Sample Frame (1-Yes; More than one per study possible)		
Newpaper Ads	17	46%
Online ads or job bank	19	51%
Other (e.g., industry lists, employment agencies)	2	5%
Gender of Applicants [(c)]		
Female	3	10%
Male	14	47%
Both	13	43%
Criminal Record		
Yes (some auditors in study have fictitious criminal records)	9	29%
Region [(d)]		
Northeast	6	19%
Midwest	6	19%
South	3	10%
West	7	23%
Occupational Categories (more than one per study possible) [(e)]		
Includes blue collar	14	47%
Includes office focus	24	80%
Includes restaurant jobs	20	67%

[(a)] Tabulated at effect level, excluding four studies with fieldwork dates before 1989

[(b)] Studies with =>2 applications/job from different racial/ethnic groups that provide counts at the pair level

[(c)] Two studies did not clearly state gender of testers. These were coded to both.

[(d)] Eight studies (not included in this tabulation) use national or multi-regional samples.

[(e)] One study lacking clear occupatonal information was coded as not including restaurant jobs

Table S2: Characteristics of Field Experiments of Discrimination Since 1989 (Continuous variables)

	Mean[a]	Std. Dev.
Year of Fieldwork [b]	2005.3	8.7
Jobs Applied For		
In-person Audits	185.7	146.6
Resume Audits	1422.3	1636.8
Applications Submitted		
In-person Audits	440.4	244.2
Resume Audits	3061.2	2331.7
Positive Response Rates of Callbacks		
White Auditors/Resumes	25.1%	18.6%
Non-white Auditors/Resumes	18.7%	15.6%
Response Ratio (white/minority)	1.42	0.41
Unemployment Rate [c]	6.6%	1.5%
Education in Years [d]	13.6	1.7

[a] Unweighted means at the effect level, n=30

[b] For studies that do not indicate year of fieldwork (n=7), we coded fieldwork year to be year of publication minus 2 years for published articles, and minus 1 year for reports and working papers.

[c] Average unemployment rate of metropolitan areas in the study. Rates are averaged across months or years of the fieldwork

[d] Nineteen effect sizes are based on applicants with a single education level, the other eleven included applicants with varying levels of education. We use a continuous education measure coded high school degree=12, some college, no degree =13, associate's degree=14, four year college degree=16, graduate degree=17.

Table S3: Random-Effects Meta-Regressions of Log Discrimination Ratios on Year

Outcome: Log of Ratio of Callback Proportions White/Minority

Predictor Variable	African-Americans Since 1989	African-Americans All Years	Latinos Since 1989
Fieldwork Year	0.004	0.007	-0.006 +
	(0.005)	(0.005)	(0.003)
Tau-squared	0.021	0.033	0.000
N (effects/studies)	21	25	9

Notes: Models are estimated with a constant, but it is not shown.
Standard errors are shown in parentheses. Significance tests employed the Knapp-Hartung modification.
Tau-squared is the estimated variation between studies (attributable to study differences).
** = p<.01; * = p<.05, + = p< .1; two-tailed tests

Table S4: Random-Effects Meta-Regressions of Log Discrimination Ratios, African-American, Studies After 1989

Outcome: Log of Ratio of Callback Proportions White/African-American

		Modified Outcomes			Base Outcome with Controls			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Predictor Variable	Base Model, No Controls	Job Offer in Place of Interview / Callback	No Criminal Record or Disability	Resume Audit Studies Only	Applicant Attributes	Metro. & Region	Occupations	Trimmed
Fieldwork Year	0.004	-0.008	0.006	-0.002	0.008	0.001	0.008	0.013
	(0.005)	(0.007)	(0.004)	(0.007)	(0.010)	(0.011)	(0.011)	(0.009)
Study Method is In-Person Audit					0.315	0.228	0.361	0.279
(1=In-person audit, ref.=resume audit)					(0.182)	(0.273)	(0.212)	(0.164)
Applicants Male Only (1=yes, ref.= male & female)					-0.007	-0.166	-0.073	-0.139
					(0.113)	(0.139)	(0.152)	(0.126)
Applicants Female Only (1=yes, ref.= male & female)					0.100	-0.311	0.151	
					(0.155)	(0.457)	(0.173)	
Applicant Education in Grades Completed					0.047	0.060	0.063	
					(0.047)	(0.067)	(0.063)	
Some Applicants have (Falsified) Criminal Records (1=yes)					0.206	0.109	0.158	-0.035
					(0.166)	(0.245)	(0.206)	(0.149)
Unemployment Rate of Metropolitan Area(s), Percentage						-0.006		
						(0.052)		

Table S4, Continued

	(1) No Controls	(3) Job Offer in Place of Interview / Callback	(4) No Criminal Record or Disability	(4) Resume Audit Studies Only	(5) Applicant Attributes	(6) Metro. & Region	(7) Occupations	(8) Trimmed
Region = Midwest (1=yes, vs. reference of multi-region)						0.405		0.147
						(0.367)		(0.113)
Region = Northeast (1=yes)						0.372 +		0.155
						(0.198)		(0.168)
Region = South (1=yes)						-0.005		
						(0.190)		
Region = West (1=yes)						0.023		
						(0.195)		
Includes blue collar occupations (1=yes)							0.229	0.117
							(0.185)	(0.158)
Includes occupations with an office focus (1=yes)							-0.009	
							(0.171)	
Includes restaurant occupations (1=yes)							-0.073	
							(0.188)	
Tau-squared	0.021	0.021	0.012	0.004	0.023	0.023	0.029	0.021
Number of Studies	21	21	21	13	21	21	21	21

Notes: Standard errors are shown in parentheses. Significance tests employ the Knapp-Hartung modification.
Models are estimated with a constant, but constant is not shown.
Tau-squared is the estimated variation between studies (attributable to study differences).
** = p<.01; * = p<.05, + = p< .1; two-tailed tests

Table S5: Random-Effects Meta-Regressions of Log Discrimination Ratios, Latinos, Modified Outcomes

Outcome: Log of Ratio of Callback Proportions White/Latino

	Latinos		
Predictor Variable	Job Offer in Place of Interview / Callback	No Criminal Record or Disability	Resume Audit Studies Only
Fieldwork Year	-0.009	-0.006	-0.008
	(0.005)	(0.004)	(0.005)
Tau-squared	0.000	0.000	0.000
N (studies)	9	9	5

Notes: Standard errors are shown in parentheses.   Significance tests employ the Knapp-Hartung modification.
Models are estimated with a constant, but constant is not shown.
Tau-squared is the estimated variation between studies (attributable to study differences).
None of the coefficeints are significant at p<.1, two-tailed tests.

Table S6: Random-Effects Meta-Regressions of Log Discrimination Ratios, African-Americans and Latino Effect Sizes Pooled

Outcome: Log of Ratio of Callback Proportions White/Minority

Predictor Variable	(1) No Controls	(2)	(3) Applicant Attributes	(4) Metro. & Region	(5) Occupations
	Since 1989	All Years	Since 1989	Since 1989	Since 1989
Fieldwork Year (Year 2015 = 0)	0.004	0.007	0.003	-0.004	0.006
	(0.005)	(0.006)	(0.010)	(0.008)	(0.011)
White/Hispanic	-0.193 +	-0.209 +	-0.200	-0.198	-0.233
(1=yes, vs. white/black)	(0.079)	(0.090)	(0.131)	(0.115)	(0.136)
White/Hispanic * Fieldwork Year	-0.009	-0.011	-0.010	-0.009	-0.009
	(0.005)	(0.007)	(0.008)	(0.007)	(0.010)
Study Method is Audit			0.1248	0.055	0.198
(1=Audit, ref.=correspondence)			0.1109	(0.108)	(0.126)
Applicants Male Only (1=yes, ref.= male & female)			-0.033	-0.147	-0.082
			(0.077)	(0.102)	(0.137)
Applicants Female Only (1=yes, ref.= male & female)			0.006	-0.491	0.075
			(0.087)	(0.318)	(0.050)
Applicant Education in Grades Completed			0.024	0.047	0.030
			(0.041)	(0.033)	(0.032)
Some Applicants have (Falsified) Criminal Records (1=yes)			0.179	0.053	0.114
			(0.194)	(0.118)	(0.243)
Unemployment Rate of Metropolitan Area(s), Percentage				0.001	
				(0.026)	

Table S6 Continued:

	(1)	(2)	(3)	(4)	(5)
	No Controls		Applicant Attributes	Metro. & Region	Occupations
	Since 1989	All Years	Since 1989	Since 1989	Since 1989
Region = Midwest (1=yes, vs. reference of multi-region)				0.534	
				(0.302)	
Region = Northeast (1=yes)				0.355 *	
				(0.133)	
Region = South (1=yes)				0.011	
				(0.194)	
Region = West (1=yes)				0.047	
				(0.109)	
Includes blue collar occupations (1=yes)					0.200
					(0.108)
Includes occupations with an office focus (1=yes)					0.041
					(0.088)
Includes restaurant occupations (1=yes)					-0.061
					(0.162)
Tau-squared	0.019	0.014	0.020	0.014	0.024
Number of Effects	30	34	30	30	30
Number of Studies	24	28	24	24	24

Notes: Standard errors are shown in parentheses. Significance tests employ the Knapp-Hartung modification.
Models are estimated with a constant, but constant is not shown.
Correlated effects model, robust standard errors with small-sample adjustment (see Tipton 2015).
Tau-squared is the estimated residual variation between studies.
** = p<.01; * = p<.05, + = p< .1; two-tailed tests

Table S7: Random-Effects Meta-Regression of Log Discrimination Ratios, African-Americans, Publication Predictors

Outcome: Log of Callback Ratio White/African-American

Predictor Variable	Coefficient (SE)
Fieldwork Year	-0.001
	(0.009)
Authors=Advocacy Groups (1=yes, vs. academic authors)	-0.130
	(0.211)
Authors=Government (1=yes, vs. academic)	0.237
	(0.357)
Publication Type is Journal (1=yes)	0.092
	(0.122)
Race is not primary focus (1=yes)	0.051
	(0.112)
Tau-squared	0.025
Number of Studies	21 (all since 1989)

Notes: Standard errors are shown in parentheses. Significance tests employ the Knapp-Hartung modification.
Models are estimated with a constant, but constant is not shown.
Tau-squared is the estimated variation between studies (attributable to study differences).
** = p<.01; * = p<.05, + = p< .1; two-tailed tests

Table S8: Random-Effects Meta-Regressions of Discrimination, Difference in Proportions Outcome

Outcome: Difference in Callback Proportion, White to Minority

Predictor Variable	(1) African-Americans		(2) Latinos	
Fieldwork Year	0.001		-0.005	**
	(0.001)		(0.001)	
Study Method is In-Person Audit (1=In-person audit, ref.=resume audit)	0.095	**		
	(0.027)			
Applicants Male Only (1=yes, ref.= male & female)	0.006			
	(0.014)			
Applicants Female Only (1=yes, ref.= male & female)	0.040			
	(0.023)			
Applicant Education in Grades Completed	0.004			
	(0.006)			
Some Applicants have (Falsified) Criminal Records (1=yes)	0.018			
	(0.023)			
Tau-squared	0.0003		0.0001	
Number of Studies	21		9	

Notes: Outcome is the difference in the proportion of callbacks received by white applicants minus the proportion received by minority applicants.

Standard errors are shown in parentheses. Significance tests employ the Knapp-Hartung modification.

Models are estimated with a constant, but constant is not shown.

Tau-squared is the estimated variation between studies (attributable to study differences).

** = p<.01; * = p<.05, + = p< .1; two-tailed tests

Table S9: Random-Effects Meta-Regressions of Discrimination, Log Odds Ratio Outcome

Outcome: Log of Odds Ratio of Callback, White to Minority

	(1)	(2)
Predictor Variable	African-Americans	Latinos
Fieldwork Year	0.005	-0.021 **
	(0.011)	(0.005)
Study Method is In-Person Audit	0.420 +	
(1=In-person audit, ref.=resume audit)	(0.202)	
Applicants Male Only (1=yes, ref.= male & female)	-0.006	
	(0.121)	
Applicants Female Only (1=yes, ref.= male & female)	0.136	
	(0.174)	
Applicant Education in Grades Completed	0.058	
	(0.051)	
Some Applicants have (Falsified) Criminal Records (1=yes)	0.265	
	(0.182)	
Tau-squared	0.0002	0.0024
Number of Studies	21	9

Note: Standard errors are shown in parentheses. Significance tests employ the Knapp-Hartung modification.
Models are estimated with a constant, but constant is not shown.
Tau-squared is the estimated variation between studies (attributable to study differences).
** = p<.01; * = p<.05, + = p< .1; two-tailed tests

Table S10: OLS Regressions of Fieldwork Year on Other Covariates, African Americans, Studies After 1989

Outcome: Year of Fieldwork of Study

Predictor Variable	Coef. (SE)
Study Method is In-Person Audit	-4.589
(1=In-person audit, ref.=resume audit)	(10.521)
Applicants Male Only (1=yes, ref.= male & female)	6.248
	(5.821)
Applicants Female Only (1=yes, ref.= male & female)	0.093
	(21.193)
Applicant Education in Grades Completed	0.805
	(2.675)
Some Applicants have (Falsified) Criminal Records (1=yes)	18.072
	(10.972)
Unemployment Rate of Metropolitan Area(s), Percentage	0.397
	(2.298)
Region = Midwest (1=yes, vs. reference of multi-region)	-1.097
	(15.232)
Region = Northeast (1=yes)	11.393
	(8.431)
Region = South (1=yes)	8.290
	(10.238)
Region = West (1=yes)	6.168
	(8.740)
Includes blue collar occupations (1=yes)	-4.680
	(9.490)
Includes occupations with an office focus (1=yes)	5.570
	(7.028)
Includes restaurant occupations (1=yes)	-14.537
	(13.880)
F-statistic (13, 7 df)	1.930
P-value for F-statistic	0.194
Number of Studies	21

Notes: Standard errors are shown in parentheses.

Models are estimated with a constant, but constant is not shown.

None of the coefficients produce significance tests with p-values < .1